

DOCUMENT RESUME

ED 282 423

FL 016 714

AUTHOR Wesdorp, Hildo
 TITLE Backwash Effects of Multiple-Choice Language Tests: Myth or Reality?
 PUB DATE May 83
 NOTE 18p.; In: van Weeren, J., Ed. Practice and Problems in Language Testing 5; see FL 016 706. Small print in figures may affect legibility.
 PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
 EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS Dutch; Elementary Secondary Education; English (Second Language); Foreign Countries; French; German; *Language Tests; *Multiple Choice Tests; Native Language Instruction; Reading Comprehension; Reading Tests; *Second Language Instruction; Student Attitudes; *Test Bias; *Testing Problems; Test Use; Uncommonly Taught Languages
 IDENTIFIERS *Comprehensive Examinations; *Netherlands

ABSTRACT

Since the mid-1960s, the use of objective testing in education, and multiple-choice tests in particular, has increased dramatically in the Netherlands. Since the founding of the Dutch National Institute for Educational Measurement in 1968, the construction of many achievement tests has been concentrated there. This general changeover from traditional to objective testing also applies to language teaching. The construction of many multiple-choice tests both for Dutch as a native language and for French, German, and English, the three most important second languages taught in Dutch schools has resulted. Multiple-choice tests for Dutch are now part of the final testing program at the end of primary education, which plays an important part in the selection process for secondary education. Some types of secondary education also incorporate multiple-choice tests for Dutch in the final examinations that determine a student's admission to higher education. The same is true for foreign language tests.
 (Author/MSE)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

1 Introduction

Since the mid 1960's, the use of objective testing in education, in particular of multiple-choice tests, has increased dramatically in the Netherlands. In 1968, the National Institute for Educational Measurement (Cito) was founded and since then the construction of many achievement tests has been concentrated there. This general changeover from traditional testing to objective testing also applies to language teaching. It has resulted in the construction of a fair number of multiple-choice tests, not only for Dutch as a mother-tongue, but also for French, German and English, the three most important foreign languages taught in Dutch schools. Multiple-choice tests for Dutch as a mother-tongue are now part of the final testing program at the end of primary education which plays an important part in the selection for secondary education. Some types of secondary education also incorporate multiple-choice tests for Dutch in the final examinations which determine a student's admission to higher education. The same holds for foreign language tests: they form an important part of the final examinations in various types of secondary education.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. van Weeren

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to improve
reproduction quality.

- Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

From an educational point of view, the introduction of a new type of test for an important selection procedure is certainly not a trivial matter: the new tests are frequently feared for their possible influence on education in general. Most of the time they provoke heated discussions in education journals, but the debate generally subsides after a while, when teachers have become used to the new tests. Eventually, the educational system accepts the new situation and the discussion stops altogether. At first glance, such a development seems perfectly acceptable, but this course of events also has its problematic sides. The developers of the new language tests are primarily interested in perfecting the technical quality of their measuring instrument and do not worry so much about its effects on the educational system. Nevertheless, these effects determine in part the validity of the newly introduced instrument. After all, tests are made to accommodate educational goals and their content-validity is defined as the relationship between the content of the test and of the curriculum. If a test with only moderate content-validity is introduced in a high-status testing situation, for example in an important selection procedure, the chances are that the educational system will adapt to the questions being asked in the test. As time goes on, this process will increase the validity of the test content and, after a while, the fit between the educational system and the test will be perfect, giving the test perfect content-validity, while the original discussion about the discrepancy between

ED282423

FL016714

test content and educational goals recedes into the background.

Those who construct tests should be more aware of such developments. Such an awareness should, in the first place, lead to the development of tests with a perfect fit to the existing educational goals, so that backwash effects will be out of the question. In the second place, this awareness should be an incentive for research into the possible effects of introducing the measuring-instrument in question, when there is doubt about the desired fit. In this way, test-constructors would achieve a more balanced view of the various aspects of their testing instruments.

In the heated debate on the effects of the introduction of the new type tests, four objections are most often heard in the Netherlands. First, discriminatory effects were expected. The new tests were thought to put certain social groups at a disadvantage. Second, an impoverishment of the curriculum was expected. It was thought that skills which could not be measured through multiple-choice tests would not be practiced in the schools and would, therefore, gradually be lost. In the third place an impoverishment of teaching methodology was expected. Supposedly, a final language examination with multiple-choice tests would preclude the use of certain teaching methods, causing a gradual decline of their use. And the fourth drawback was the possible influence of multiple-choice tests on the study-habits of students. Students were expected to prepare themselves in a different way when used to multiple-choice tests. In addition, certain emotional aspects of multiple-choice tests (e.g. fear of test-taking) were mentioned.

This paper reports on various research-projects carried out during the years 1976-1979. They were all aimed at the question: can we find some of the mentioned backwash-effects? Research on the four groups of effects, mentioned above, will be reported in the sections 2-5.

From earlier surveys and from the Dutch literature on the subject, it became clear that to many people the introduction of multiple-choice tests in examinations implies an increase in selection on the basis of socio-economic status. Multiple-choice tests are looked upon as undemocratic and are thought to place certain social groups or regions at a disadvantage when such tests are part of an examination.

These opinions often appear to be based on the idea that multiple-choice tests appeal more to certain verbal skills than traditional instruments. This is seen to be a disadvantage for certain social groups, particularly when they are also confronted with middle-class language, of which they do not have a good command.

In this investigation the supposedly discriminatory effect of multiple-choice tests has been assessed in two different ways. One part of the investigation was carried out in the final year of HAVO schools, for English, while the other part was directed at the LBO examinations for Dutch*. The object of the investigation was to find out whether multiple-choice tests discriminate against certain social groups, as compared to traditional tests.

For the first investigation, parallel instruments were constructed which were as much equivalent as possible. One consisted of multiple-choice questions and one of essay questions. Two versions were made of this instrument, so that half of the students were examined first on the multiple-choice questions and then on the essay-questions, while the other half were tested in the opposite order. As far as content was concerned, all students answered the same sixteen questions. The tests, together with questions concerning the occupation of the student's father, were administered to 412 final year students from twelve HAVO-schools.

The average degree of difficulty (p-value) of the multiple-choice questions was .57 and of the open questions .34. The test which first presented the multiple-choice questions, followed by the essay-questions, showed slightly better results than the test presenting the questions in the opposite order. The reliability (Cronbach's α) of both test versions was .63. The average scores for students of different socio-economic status is shown in Table 1. Considering the low reliability of the test, the differences in the scores of Table 1 should be interpreted with caution.

* HAVO: middle stream of secondary education, 5 years.

LBO : lower stream of secondary education (vocational track), 4 years.

Table 1 Mean test-scores HAVO-English

Socio-economic status	N	Multiple-Choice	Open-Q	Difference-scores
1 unskilled and skilled workers	64	4.74	2.40	2.34
2 lower level employees	108	4.43	2.84	1.59
3 small business-owners	45	4.11	2.28	1.83
4 mid-level employees	44	4.68	3.07	1.61
5 high level employees	69	4.46	2.59	1.87

The test results have been analyzed in a number of ways. Separate analyses of variance for both test halves did not show any interaction effect between socio-economic status and question format. An analysis of score differences did not indicate a discriminatory effect for multiple-choice questions, nor did the relative results for boys and girls on essay and multiple-choice questions yield any difference. Significant differences were found between schools, but there is no question of an interaction effect between school and socio-economic status; in other words, none of the schools was found to have a socially discriminating influence.

The results of this first investigation do not in any way indicate an 'anti-democratic effect' for multiple-choice questions. It is, of course, possible that the attempt to assess the effect of the differences in question format as accurately as possible, by means of keeping the question-content constant, has removed the essential differences between the two types of questions from the experiments. The presumed differences between multiple-choice and open questions are not limited to the presence or absence of a list of alternative answers under the question. One criticism often levelled against objective tests is their poorer content as compared to traditional tests.

A comparison of traditional and multiple-choice tests - both representative for their category - was the purpose of the second investigation. The current situation for LBO final exams offered a good opportunity for such a comparison. One track of LBO has a two-part final examination for a number of subjects, including Dutch. The students take one exam, consisting of open questions and one consisting of multiple-choice questions. By collecting both types of results for each candidate, as well as data on socio-economic status, a comparison could be made of the difference in achievement on either type of test for students from various social groups. By using data from actual exams, maximum student motivation and achievement is guaranteed, while artifacts of an experimental situation can also be avoided.

For 258 final-year students from seven LBO-schools the final exam results for Dutch were collected in 1977 and a classification was made of the fathers' occupations. In Table 2 the exam results are given as follows:

- Average scores and standard deviations for the essay-questions on the examinations: maximum score 57.
- Average scores and standard deviations for the

multiple-choice part of the examination: maximum score 50.

Table 2 Mean scores and standard deviations for students of different socio-economic status: final examination Dutch (LTO 177)

	N	Multiple-Choice		Open Questions	
		M	SD	M	SD
1 Unskilled labor	29	26.69	4.67	36.69	6.70
2 Skilled labor	95	28.74	5.27	36.18	6.62
3 Lower level employees	46	28.48	5.51	37.02	5.69
4 Small business-owners	58	29.52	4.85	36.81	6.68
5 Mid-level and high level empl.	30	29.83	5.09	36.43	7.71

The table shows that the students scored higher on the essay questions than on the multiple-choice part of the test, while the standard deviations on the essay questions are also higher. The correlation of .36 between the two formats is significant, but rather low, presumably because both test forms measure a number of skills simultaneously. Neither the open questions nor the multiple-choice questions showed significant differences, according to socio-economic status, although the averages for multiple-choice questions for higher socio-economic status students were somewhat higher. The average differences of the five status groups seem to show the predicted tendency, but there was no question of any significant difference. Since the reliability of the score differences is probably also quite low in this case, given the low reliability of the scores for both question formats, these results should be interpreted with caution. By means of an analysis of variance, a significant difference between the seven schools on both types of questions could be found, but there was no indication of an interaction effect between school and socio-economic status. In other words, none of the schools seemed to show an influence of socio-economic status on the differences in results for the two test formats.

From this investigation it may be concluded that a socially discriminating effect for multiple-choice tests in the 1977 final examination for Dutch in LBO-schools and in the constructed MAVO-tests for English could not be demonstrated. Achievements on multiple-choice and open question-ended tests for Dutch and English were not found to be different for the various participating social groups.

In a survey among foreign language teachers about their attitudes towards multiple-choice tests we found that considerable minorities of teachers believe in the discriminatory effects of the multiple-choice reading comprehension tests, that are part of the final examinations. A statement like 'Reading comprehension tests (in multiple-choice format) are undemocratic; certain social classes or regional groups are put at a disadvantage' was supported by 20% of the teachers. And a statement like 'Because of the introduction of multiple-choice reading comprehension tests in the examinations the selection according to social class has been intensified' was supported by 18% of the teachers.

Although our investigations did not find any proof of such socially discriminating effects, important minorities of teachers still strongly believe the opposite. Research findings and teacher opinions are conflicting.

3.1 The influence of multiple-choice testing on the writing skills of students

In 1966, the first achievement test was administered in Amsterdam in the final year of primary education. This achievement test has slowly acquired a certain degree of status, because its results determine in part in which type of secondary education the student will be placed. The instrument consists entirely of multiple-choice tests on all the important subjects taught in primary school. For Dutch as a mother tongue the test contains subtests on reading comprehension, spelling, grammar, and an attempt to measure writing skills in an indirect way. An essay containing various mistakes is used to assess a number of writing skills. The student is requested to correct the mistakes, choosing one from a few alternatives. The mistakes concern skills in a variety of areas. These areas are: writing conventions, i.e. dealing with problems of spelling and punctuation; syntax, appropriate and effective language-use, i.e. stylistic nuances; organization of the essay, i.e. dealing with structural mistakes, such as lack of necessary information, poor use of paragraphs; and content, i.e. mistakes in the content of the essay. The first year in which the exam was given (1966), all participants also wrote an essay, which was not evaluated in any way.

It was thought that the limited scope of the test content (only multiple-choice items, no essay questions, no testing of productive language skills) would lead to less attention being paid to independent essay-writing, eventually leading to a decrease in writing skills on the part of the students. The 1966 requirement to write an additional essay made it possible to compare writing skills in 1978 and 1966. A comparable sample of students was required to write an essay, similar to the one in 1966. The two sets of essays were then compared for quality, presupposing that the effect of twelve years of multiple-choice testing would have resulted in poorer quality of the 1978 essays.

Two comparisons were made in this investigation. First, 64 'old' and 64 'new' essays were compared for overall quality by 15 judges, giving a global judgement. Obviously, there were no differences in format between the old and the new essays and they were mixed randomly. Secondly, 32 old and 32 new essays were judged analytically: i.e. judgements on 'content', 'style', and 'writing conventions' were elicited from 16 (different) judges. Differences were expected, both in the comparison of the global judgements and of the analytic judgements, which could presumably be explained by the 'impoverishment-effect' of the achievement test. The

reliability of the jury-judgements (the degree of consensus with other juries of equal size and competence) varied from .78 to .95 for the various criteria. Table 3 shows some results of the investigation. The global judgement the total analytic judgements were given on a ten-point scale, the type of scale generally used in schools in the Netherlands, while the analytic categories were judged on a four-point scale.

Table 3 Mean scores and standard deviations for old (1966) and new (1977) essays, by means of global and analytic judgement by juries of 15 and 16 members, respectively

	GLOBAL		ANALYTIC							
	Old	New	Total		Content		Style		Conventions	
	Old	New	Old	New	Old	New	Old	New	Old	New
M	6.50	6.60	5.96	6.04	2.64	2.77	2.84	2.79	2.95	2.80
SD	0.68	0.67	0.67	1.02	0.38	0.56	0.44	0.57	0.52	0.61
n	64	64	32	32	32	32	32	32	32	32

None of the expected differences was actually found in the investigation, leading to the conclusion that no impoverishment as a result of the use of multiple-choice achievement tests is to be found in writing skills of the pupils. Enrichment in the more technical aspects of writing skills, such as 'writing conventions' could not be found either, even though the test does especially stress such points every year.

3.2 A comparison of old and new reading comprehension exams for the mother-tongue at the end of secondary education

For one type of secondary education, Lower Technical Schools (LTO), it was possible to compare traditional reading comprehension exams using a text with open questions, with the new form of reading comprehension tests using a text followed by a number of multiple-choice questions. For a number of years, these two examination formats existed side by side, which made it possible to compare the skills that were being measured by the two types of tests between 1971 and 1977. This was done by using the criteria for measuring reading comprehension skills as formulated by F.B. Davis in: The literature of research in reading with emphasis on models. Rutgers, The State University, New Brunswick (N.J.), 1971

Nine operational skills of comprehension in reading among mature readers

- 1 Word knowledge, as measured by recognition vocabulary items
- 2 Ability to select the appropriate meaning for a word or phrase in the light of its particular contextual setting

- 3 Ability to follow the organization of a passage and to identify antecedents and find references in the passage
- 4 Ability to select the main thought of a passage
- 5 Ability to answer questions which are answered directly in the passage
- 6 Ability to answer questions which are answered in the passage but not in the words in which the question is asked
- 7 Ability to draw inferences from the passage about the content of the passage
- 8 Ability to recognize the literary devices used in a passage and to apprehend its tone and mood
- 9 Ability to determine the writer's purpose, intent and point of view; i.e., to draw inferences about the author

Table 4 gives the percentage for each category of the Davis system which is required for answering the two types of exam question, according to the judgements of a three-member jury. The open question exams contained a total of 276 questions, the multiple-choice exams contained 447 questions. The average inter-judge reliability was 57%, which is why the results for the three judges have not been collapsed into one figure. According to all three judges, there are differences in the way the skills for the two types of examinations are divided among the Davis categories.

Table 4 Classification of traditional and objective examinations for Dutch reading comprehension for LTO-schools from 1971 to 1977, by three judges, according to the Davis criteria (in percentages)

Examination	Categories									Not
Open Question	1	2	3	4	5	6	7	8	9	classified
judge I	20	7	20	3	19	17	8	6	0	0
judge II	12	13	16	0	31	10	6	2	0	9
judge III	24	11	19	0	32	6	5	0	0	4
Multiple-choice										
judge I	6	11	6	10	23	26	9	6	3	0
judge II	1	18	7	3	31	16	18	4	3	0
judge III	4	16	10	4	27	20	13	4	2	1

The three judges found that on open question exams, questions about word knowledge (Davis skill 1), and grammatical structure and function (skill 3) are asked more frequently. Both multiple-choice and open questions exams frequently demand answers that are found directly in the passage (skill 5). Multiple-choice questions often require the ability to select appropriate meaning from context, recognize paraphrases, or draw inferences (skills, 2, 4, 6, 7 and 8). In order to see to what extent these differences are due to a general change in exams over the period from 1971 to 1977, it was decided to compare the Davis classification for the last three

exams separately (two exams from 1976 and one from 1977). Table 5 shows the required skills in percentages, according to the judgements from three judges.

Table 5 Classification of traditional and objective examinations for Dutch reading comprehension for 1976 and 1977 by three judges, according to the Davis criteria (in percentages)

Examination	Categories								
Open Questions	1	2	3	4	5	6	7	8	9
judge I	1	3	10	5	41	27	10	2	1
judge II	0	2	13	0	62	15	7	1	1
judge III	0	10	13	0	65	10	1	0	0
Multiple-choice									
judge I	4	9	3	11	25	26	8	9	5
judge II	0	16	5	5	30	21	13	4	5
judge III	4	13	7	6	29	20	11	5	4

Testing the frequency-distribution of the various skills for the two exam types yielded significant values for all three judges. According to all three judges, skills 6, 7, 8 and 9 were required more in multiple-choice tests. This means that, according to the judges, the exams in the last two years required more ability to paraphrase, draw inferences, recognize literary devices, and determine the writer's intent and point of view. In 1976 and 1977, both multiple-choice and open-question exams required less word knowledge than did earlier exams.

The conclusion that may be drawn from this investigation is that the analyzed LTO-reading comprehension examinations have not become poorer in content as a result of the introduction of multiple-choice tests. On the contrary: a certain degree of enrichment can be found in comparison with the open question examinations. Inasmuch as the differences between the two types of exams have become smaller in recent years, or have disappeared completely, this seems to have resulted from an improvement in the essay questions after the introduction of multiple-choice questions in final exams.

In a survey among foreign language teachers we also asked for their opinions on the subject of the impoverishment of the curriculum, especially the foreign language curriculum in the various streams of our Dutch secondary system. The teachers turned out to be rather critical. Important minorities supported the following statements: The foreign language reading comprehension tests put too much stress on general reasoning questions, whilst the knowledge of the second language is of secondary importance (67%).

The foreign language reading comprehension test neglect:

- lyrical texts, emotion-evoking textelements (66%)
- reacting to texts, criticizing the texts, text analysis, evaluating the texts (81%)
- humorous, ironical, witty texts (48%).

Our two empirical investigations were both aimed at the discovery of effects on the mother-tongue curriculum. We could not find any evidence of the existence of such effects. Our foreign language teachers however, are quite sure that in their field the final examinations in multiple-choice format lack validity and therefore must have certain backwash effects.

4.1 The influence of 11+ tests on mother-tongue teaching in the primary schools

Another investigation was designed to see whether the achievement test mentioned in 3.1, with its limited coverage of the goals of mother-tongue education, had any effect on teaching practices. To this end, a comparison was made of the time spent on various parts of the curriculum and of types of teaching in 102 schools. 51 schools had been using the test for quite some time and 51 had not used the test at all. Obviously, the expected differences were that the 'test schools' would pay more attention to components of the curriculum that are part of the test and less to components that will not be tested, in comparison with the 'non-test schools'. Two types of information were elicited by means of a questionnaire: an assessment of the time spent on various types of teaching and of the teacher's attention to different parts of the language curriculum.

Table 6 Time spent on various teaching methods in 51 'test-schools' and in 51 non-test schools' a)

<u>type of teaching</u>	<u>test schools</u>		<u>non-test schools</u>	
	n	med.	n	med
teaching in class	47	3.86	50	3.25
teaching individually	48	2.07	50	2.17
teaching small groups	47	2.00	50	1.62
<u>teacher activities</u>				
teaching	44	2.21	41	0.83
narrating	50	1.63	48	1.29
explaining	51	1.74	48	1.90
demonstrating	49	0.98	45	1.06
individual activity	50	2.17	47	2.00
class discussion	51	1.07	49	1.16
oral quizzing	48	0.95	46	0.67
written quizzing	50	1.07	48	0.90

a) Time spent is expressed on a seven-point scale, ranging from 1 (almost never) to 7 (more than twenty hours per week).

Table 6 shows that few differences could be found. The differences were subjected to the Mann-Whitney U-test, which showed that both oral and written quizzing is done more frequently in the 'test schools'. Table 7 shows the way in which the teacher's attention was divided among the various parts of the language curriculum.

Table 7 Time spent on language curriculum components in
51 test-schools and non-test schools b)

Curriculum component	Part of test?	Test schools		Non-test schools	
		n	med.	n	med.
Technical reading	no	51	1.00	50	1.07
Reading comprehension	yes	51	1.36	50	1.23
Essay writing	no	51	0.86	49	0.84
Grammar	optional	48	0.97	48	1.01
Spelling	yes	51	1.25	49	1.13
Style exercises	yes	49	1.14	49	1.16
Listening comprehension	no	49	0.97	48	0.97
Speaking	no	50	0.95	50	1.00

No differences could be traced in the time spent by teachers on various parts of the curriculum, even though some differences had been expected. Consequently, the investigation led to the conclusion that the use of achievement tests has no clear influence on teaching practices.

4.2 The use of multiple-choice tests (as teaching aids) in secondary education

The degree of test-use in the classroom could be considered as an indicator of possible influence of multiple-choice final examinations on the curriculum. The more frequently multiple-choice tests are used in class, the more likely it is that teaching-content is reduced to teaching only what will be tested in the final examination. In this section a summary is given of a survey of the frequency with which multiple-choice tests are used for teaching in the final years of secondary education, as well as the other uses to which such tests are put.

A questionnaire was sent to a sample of schools for secondary education, in particular to the departments whose subjects are partially or completely tested by means of a multiple-choice test in the written final examination. 965 schools were approached, comprising 4471 departments. The percentage of responses was 47% (2102 questionnaires). The survey yielded a large amount of interesting data, which cannot all be summarized here, but the most striking results for language teaching will be given below.

The first part of the survey was concerned with the format of test questions used in the classroom. In the final year, the percentage of multiple-choice questions compared to essay and short-answer questions was 21% and 15% respectively for French, 23% and 11% for German, and 27% and 14% for English. For Dutch the percentages were higher: 34% compared to 30%. These figures imply that multiple-choice tests only have limited influence on classroom test format, while the influence is greatest for Dutch.

The second part of the survey was designed to find out

whether multiple-choice tests are used as teaching aids in the classroom and as homework assignments. The results for language teaching have been summarized in Table 8 and show that multiple-choice tests are not being used on a large scale. Special exercises to train students for multiple-choice questions are given frequently, though.

Table 8 Use of multiple-choice questions in the classroom and as homework-assignment in the two final years of secondary school, in percentages

	Dutch	French	German	English
Multiple-choice questions used as teaching aid in teaching subject material	24	20	19	25
Students given multiple-choice questions in class for practising subject material	52	34	31	38
Students given multiple-choice questions in class for practising multiple-choice questions	58	40	54	51
Students given multiple-choice questions as homework for practising subject material	27	24	21	27
Students given multiple-choice questions as homework for practising multiple-choice questions	31	22	29	29

In the same survey among foreign language teachers mentioned before we also asked teachers about their opinions on the influences of multiple-choice testing on foreign language teaching methodology. Most teachers are rather critical: majorities support statements like the following:

- Reading comprehension tests occupy too important a position as didactic aids (68%).
- I am speaking much more Dutch now than the language I teach (54%).
- Testing the various skills separately leads to teaching them separately (58%).
- Because of the final examinations in multiple-choice format teachers are forced to use a traditional 'frontal' class-teaching approach (59%).

But on the other hand foreign language teachers seem not so very impressed by the impact of the reading comprehension tests: 62% still think that there is only a slight chance that teaching methodology will become rigid and uniform.

Nevertheless, we still find the same phenomenon: while our empirical research cannot really prove that there are effects on teaching methodology, the teachers are quite sure they exist. The conflict between research findings and teacher opinions arises again.

Many kinds of influence on education from tests and examinations have been studied, but the influence of tests on student's study methods has been studied more than most. In a large body of investigations students have answered questionnaires stating that they study differently for multiple-choice tests than they do for traditional - open question - exams. Most of these investigations have been directed at factual school-subjects. An earlier Dutch survey among foreign language teachers showed that 43% of the teachers perceive student's study habits for foreign languages as being influenced unfavourably by the use of multiple-choice reading comprehension tests, both in exams and in the classroom.

In order to study this problem, an inventory was made in 10 Havo final-year classes of the study methods used by students who are either frequently confronted with multiple-choice tests, or with traditional testing instruments. The teachers in the participating schools had earlier been asked, how many times per month and per year they gave their students multiple-choice tests and how often open question tests. The frequencies reported by the teachers were used to form two groups of students for each subject, one with relatively little and one with relatively substantial experience with multiple-choice tests. After taking the tests, students were given a questionnaire about their preparation methods for open question or multiple-choice tests for English, depending on their reported experience with multiple-choice tests. Students who did not do any preparation at all did not fill out the questionnaire. 99 students from 5 schools were given a questionnaire about multiple-choice tests, while 85 students from another 5 schools filled out a questionnaire about preparing for open-question tests. The questionnaires contained 11 questions about preparation methods for tests, one question about student preference, and one question about the type of test-questions students found most difficult.

The experience students claimed they had had with open question tests did not always coincide with the frequency reported by the teachers. To a lesser extent, this was also the case with multiple-choice tests. Most students report little experience with open question tests: only 17% report more than one such test per year. Students appear to do little or no preparatory work for English tests. For a multiple-choice test, 2% of the students study 'always' or 'most of the time', 20% 'sometimes', and 78% '(almost) never'. For an open question test, these percentages are 32%, 20% and 48%, respectively. 84% of the students spend an average of less than half an hour preparing for a multiple-choice test, while for an open-question test 46% prepare less than half an hour, 29% between half an hour

and an hour, and 24% more than an hour. The preparation time for open question tests is mostly spent on learning words and expressions and looking through lists of idioms. Old reading comprehension exercises and tests are less popular as study material. The answers showed a strong student preference for multiple-choice tests (72%).

Although students still study a little longer for open-question tests, they generally do not prepare much at all for English reading comprehension tests. Within the group with more multiple-choice experience, a strong correlation was found between the frequency of being tested on multiple-choice tests and the degree of preparation.

In a questionnaire presented to samples of approximately 200 students for the subjects Dutch, English, and German, we also asked some questions related to the subject of this section. The following seems to be of importance:

Multiple-choice tests and the student's self-image.
Is taking multiple-choice tests the best way for students to gain insight into their own capabilities? Students do not express an opinion about this aspect. About half of the students recognize the positive feedback qualities of multiple-choice tests (for Dutch and English). For German they are more negative on this score: only a minority believe that taking multiple-choice tests tells them anything about their own level of knowledge.

Multiple-choice tests and fear of test-taking. A certain degree of test-taking anxiety is certainly produced by multiple-choice tests: 70% of the students disagree with the statement that multiple-choice tests 'let you know where you stand'. In general, they find multiple-choice tests full of treacherous pitfalls, but only a minority find multiple-choice questions so confusing as to lower the student's performance.

In our survey among foreign language teachers (mentioned above), we also asked them about their opinions on the effects of multiple-choice testing on the behaviour of the students. Large minorities were rather critical: they supported the following statements:

- The thinking pattern of the students will be unfavourably affected (39%).
- Reading comprehension tests encourage a 'fragmented' style of reading, and require too much attention to detail (50%).
- The study behaviour of pupils is disadvantageously influenced (43%).
- We teach our students to become superficial, shallow readers (19%).

Here we also find a conflict between teacher opinions and other research findings. If we believe what teachers think and say, the multiple-choice tests have rather harmful effects, especially on the ultimate goals of education: the way pupils read a text. Our other research however does not support that opinion.

On the basis of the research reported here, the overall conclusion may be drawn that in the Netherlands many of the initial fears about multiple-choice testing have turned out not to be based on facts. Although the fears among teachers more or less remain, we can only conclude that the so-called backwash effects are a myth. If they do exist, they must be so weak or small that our research methods cannot detect them. The Dutch system may have struck the proper balance between multiple-choice and traditional testing, by continuing the use of other testing techniques side by side with multiple-choice tests. This has minimized the risk of educational 'impoverishment'. It is still necessary, though, to carefully follow the developments in this area, in order to prevent an adaptation of the educational system to the testing technique.